Japanese Kokai Patent Application No. Hei 10[1998]-257436

Job No.: 228-123906 Ref.: ORDER NO. 9156
Translated from Japanese by the McElroy Translation Company
800-531-9977 customerservice@mcelroytranslation.com

(19) JAPANESE PATENT OFFICE (JP)	(12) KOKAI TOKUHYO PATENT GAZETTE (A)	(11) PATENT APPLICATION PUBLICATION NO. Hei 10[1998]-257436 (43) Publication Date: September 25, 1998

(51) Int. Cl.^{6}:		Identification Code:	FI		
H 04 N	5/93		H 04 N	5/93	Z
	5/92			5/92	H
	7/32			7/137	Z

Examination Request: Not filed　　No. of Claims: 7 (Total of 14 pages; OL)

(21) Filing No.: Hei 9[1997]-55340

(22) Filing Date: March 10, 1997

(71) Applicant: 391023987
Atsushi Matsushita
36 Kikuicho, Shinjuku-ku
Tokyo

(71) Applicant: 392008231
Kenichi Okada
4-25-12 Hongo, Bunkyo-ku
Tokyo

(72) Inventor: Atsushi Matsushita
36 Kikuicho, Shinjuku-ku
Tokyo

(72) Inventor: Kenichi Okada
4-25-12 Hongo, Bunkyo-ku
Tokyo

(74) Agent: Masatsugu Suzuki, patent attorney

(54) [Title] Automatic hierarchical structuring method for a moving picture and browsing method using same

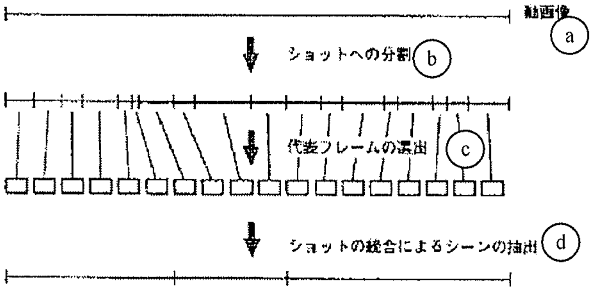
(57) Abstract

Problem

The objective of the present invention is to automatically hierarchically structure an encoded moving picture, and to obtain a video browser based on this moving picture and analytical data thereof.

Means to solve

An automatic hierarchical structuring method characterized in that a moving picture is encoded, said encoded moving picture is separated into individual shots, then the separated shots are integrated using the degree of similarity per shot, thus, extracting a scene. A moving picture browsing method characterized in that a moving picture is encoded, said encoded moving picture is separated into each shot, then the separated shots are integrated using the degree of similarity per shot, thus, extracting a scene and, thus, automatically hierarchically structuring the moving picture; this hierarchically structured data is used to comprehend the content of the entire moving picture, and the detection of a desired scene or shot is facilitated.



Key:
a　Moving picture
b　Separation into shots
c　Selection of reference frames
d　Extraction of scenes based on integration of shots

Claims

1. An automatic hierarchical structuring method for a moving picture characterized in that an encoded moving picture is separated into individual shots, and the separated shots are then integrated using the degree of similarity per shot, thereby extracting a scene.

2. The automatic hierarchical structuring method for a moving picture of Claim 1, characterized in that the moving picture is encoded by means of MPEG.

3. The automatic hierarchical structuring method for a moving picture of Claim 1, characterized in that, when a shot is detected from the encoded moving picture, it is processed at high speed using the MPEG features.

4. The automatic hierarchical structuring method for a moving picture of Claim 1, characterized in that, when the degree of similarity between shots is compared, a representative frame is extracted.

5. The automatic hierarchical structuring method for a moving picture of Claim 1, characterized in that the degree of similarity between shots is determined by means of fuzzy logic.

6. The automatic hierarchical structuring method for a moving picture of Claim 1, characterized in that the scene extraction process is determined by means of the degree of connection between defined shots.

7. A moving picture browsing method characterized in that an encoded moving picture is separated into each shot, then the separated shots are integrated using the degree of similarity per shot, thereby extracting a scene and, thus, automatically hierarchically structuring the moving picture; this hierarchically structured data is used to comprehend the content of the entire moving picture, and the detection of a desired scene or shot is facilitated.

Detailed explanation of the invention

[0001]

Technical field of the invention

The present invention pertains to an automatic hierarchical structuring method, the objective of which is to obtain a video browser that automatically hierarchically structures encoded moving pictures and operates based on this moving picture and analytical data thereof; it pertains to a browsing method using same.

[0002]

Prior art

Current moving picture information is limited to the area of simple video reproduction; in other words, moving pictures are picked up in units of frames, and when a code is appended at the time the picture is taken, it can be extracted by means of said code. In addition, reproduction time is a factor, so that when a desired frame is detected or reproduced, the frame in question can be extracted only if special relationships are clear.

[0003]

Problem to be solved by the invention

However, if certain codes are not appended, it will be extremely difficult to extract the desired frame from the provided moving picture information; if there is a time constraint, extraction becomes impossible. For example, typically, the time for one frame is 1/30 of a second, so that in one minute there are 1,800 frames, and in one hour, 108,000 frames.

[0004]

Accordingly, there is a problem in that it is not possible to extract a given frame quickly based on the aforementioned typical conventional moving picture information.

[0005]

Means to solve the problem

The present invention solves the aforementioned problem of the prior art by separating an encoded moving picture into individual shots, then integrating the separated shots using the degree of similarity per shot, thereby extracting a scene and automatically hierarchically structuring said picture; by using said data to create a moving picture browsing tool.

[0006]

In other words, the present invention concerns an automatic hierarchical structuring method for a moving picture characterized in that an encoded moving picture is separated into individual shots, and the separated shots are then integrated using the degree of similarity per shot, thereby extracting a scene; it is further characterized in that the moving picture is encoded by means of MPEG. Furthermore, it is characterized in that, when a shot is detected from the encoded moving picture, it is processed at high speed using MPEG features and in that, when the degree of similarity between shots is compared, a representative frame is extracted. It is further characterized in that the degree of similarity between shots is determined by means of fuzzy logic, and in that the scene extraction process is determined by means of the degree of connection between defined shots. Furthermore, another invention concerns a moving picture browsing method characterized in that an encoded moving picture is separated into individual shots, and the separated shots are then integrated using the degree of similarity per shot, thereby extracting a scene and, thus, automatically hierarchically structuring the moving picture; this hierarchically structured data is used to comprehend the content of the entire moving picture, and the detection of the desired scene or shot is facilitated.

[0007]

The aforementioned encoding is an MPEG-1 compression algorithm. The official name for MPEG-1 stands for "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s."

[0008]

The aforementioned hybrid encoding is performed by means of DCT and quantization, motion-compensated inter-frame prediction, and entropy encoding; however, since each of the aforementioned methods is publicly known, a detailed explanation is omitted.

[0009]

Next, an MPEG-1 encoding/decoding system will be explained using Figure 1. A video input passes through pre-processing and is input to a video encoder, undergoes system multiplexing, is input to a storage medium, and undergoes system demultiplexing, after which it is input to a video decoder, undergoes pre-processing, and is output as video.

[0010]

In Figure 2, an input picture is processed into data at the aforementioned MPEG-1 video decoder. In addition, as shown in Figure 3, the MPEG-1 video decoder processes data from an input buffer to a display buffer.

[0011]

As explained previously, MPEG-1 is intended for use in a storage medium such as a CD-ROM. The storage medium requires various special modes, such as fast-forward, rewind, random-access playback, and reverse playback. To achieve such special modes, MPEG-1 employs a Group of Pictures (hereinafter, GOP) structure.

[0012]

With MPEG-1, encoded picture data is created together with the previous and following pictures, so that complete information is not obtained with only one screen. Therefore, the data for a number of screens is compiled into the GOP unit, and random access is enabled. In other words, at least one screen included in the GOP must include screen data enclosed within only one screen (an I-picture), without using information from surrounding screens, and based on this data, the other screen data in the GOP can be reproduced. In addition, commonly one GOP consists of a group of approximately 15 pictures (Figure 4).

[0013]

With MPEG-1, both forward prediction from a past reproduction picture and backward prediction from a future reproduction picture are performed; together this is known as "bidirectional prediction."

[0014]

With MPEG-1 three types of pictures are defined to implement bidirectional prediction: I-pictures, P-pictures, and B-pictures.

[0015]

In addition, D-pictures (DC encoded images) have been defined; these are encoded using only the information within a frame and are comprised of only the DC component within a DCT coefficient. They do not exist together with the other three types of pictures in the same sequence.

[0016]

With MPEG-1, the introduction of the B-picture, which performs bidirectional encoding, significantly improves prediction efficiency, which helps to improve picture quality when high compression is used.

[0017]

As shown in Figure 6, picture data is comprised of a hierarchical structure of six layers: sequence, GOP, picture, slice, macroblock (MB), and block.

[0018]

With the aforementioned sequence layer, a bit stream representing a series of images begins with a sequence header and is followed by one or more GOPs, and ends with one sequence end code. A sequence header can be placed immediately before any GOP, but for a sequence header within a series of images, all of the data elements other than the quantization matrix must be identical to that of the initial sequence header.

[0019]

Thus, random access to anywhere within a sequence is enabled.

[0020]

In addition, The GOP layer contains a single GOP.

[0021]

Next, the picture layer contains one I-picture, P-picture, B-picture, or D-picture.

[0022]

With the slice layer, a slice is a collection of a certain number of macroblocks of a picture when raster scanning from the upper left to the lower right. Slices cannot overlap, and there can be no space between them, but the position of the slice can differ for each picture. A synchronization signal is assigned to the beginning of the slice data, which is advantageous in that, when there is a read error during decoding, synchronization can be recovered with the next slice. In addition, decoding of slice data is performed separately for each slice, so that parallel processing in units of slices to enable high-speed decoding is possible.

[0023]

In the macroblock layer, a macroblock is comprised of a 16-pixel x 16-line luminance component and two 8-pixel x 8-line color difference components that correspond to spatial positions in the picture. A single macroblock is comprised of four luminance blocks and two color difference blocks. Figure 6 shows the sequence and position of the blocks in a macroblock. Motion compensation and inter-frame prediction are performed in units of macroblocks.

[0024]

Furthermore, the block layer is the DCT processing unit and is comprised of an 8-pixel x 8-line luminance component or color difference component.

[0025]

Embodiment of the invention

The present invention is an automatic hierarchical structuring method for moving pictures whereby encoded moving pictures are separated into individual shots, and the shots are integrated using the degree of similarity per shot, thus, extracting a scene.

[0026]

Furthermore, the present invention concerns a browsing method whereby hierarchically structured data is used to comprehend the content of an entire moving picture, and the detection of a desired scene, shot, or frame is facilitated.

[0027]

By means of the present invention, the content of a moving picture can be comprehended very easily, and a prescribed location can be searched for easily.

[0028]

Application example

Next, an application example of the present invention will be explained with reference to the figures.

[0029]

First, a moving picture is separated into shots, which can be extracted by means of the physical amount of characteristic and which are relatively easy to detect. The shots are then integrated by means of the degree of similarity between shots, and a scene is extracted. In this case, it is difficult to handle a moving picture "shot," so that a number of representative frames are selected from a shot (Figure 7).

[0030]

Typically, a moving picture has a large amount of data, so that a significant amount of processing is required. In addition, for an encoded moving picture decoding is required, which further increases the amount of processing.

[0031]

With the present invention, an MPEG-1 moving picture is not completely decoded; instead, only the minimum necessary information is decoded, thus, enabling high-speed processing. Therefore, the amount of processing is reduced by using the MPEG-1 encoding algorithm characteristic, such as inter-frame prediction or the DC component, to obtain a simple picture. Accordingly, the simple decoding of an I-picture in an MPEG moving picture and frame comparison methods will be discussed, and each subsequent process will be explained in detail.

[0032]

A moving picture is comprised of several still images (frames); accordingly, the picture information for each frame is essential to the analysis of a moving picture. However, because the decoding of an MPEG-1 moving picture involves a relatively large amount of processing, it is difficult to achieve high-speed processing.

[0033]

Therefore, not all of the frames are decoded; only the I-pictures. In addition, they are not completely decoded; instead, a compressed image of a source frame is obtained by simple decoding. This simple decoding utilizes the fact that the average color for the original block is obtained when the DC component of the DCT coefficient is decoded. In other words, only the DC component of the DCT coefficient of each block is decoded, and a picture that represents each block is created using the obtained average color (Figure 8).

[0034]

The picture thus obtained will be called a "DC picture." Since the size of each block is 8x8, a DC picture is 1/8 of the vertical and the horizontal size of the original picture.

[0035]

The decoding of a B-picture or a P-picture requires the decoding of not only the picture itself but the decoding of other indirectly used information, such as motion vector information and the picture being referenced; however, an I-picture is encoded based on only what is within the frame, so that the decoding of such information is not necessary. In addition, rather than decoding the DC component of the Intra macroblock within an I-picture by means of IDCT, which involves a large number of calculations, it can be decoded by the following Equation (1). Accordingly, the DC picture can be obtained very quickly.

[0036]

Equation 1

$$\begin{aligned} Y_k &= (DY_k)/8 \\ Cb_{k'} &= (DCb_{k'})/8 \\ Cr_{k'} &= (DCr_{k'})/8 \end{aligned} \qquad (1)$$

[0037]

Here, Y_k, $Cb_{k''}$ and $Cr_{k''}$ are the luminance of the average color of each block ($_k$ and $_{k'}$ are the block numbers), and DY_k, $DCb_{k''}$ and DCr are the DC components of each block.

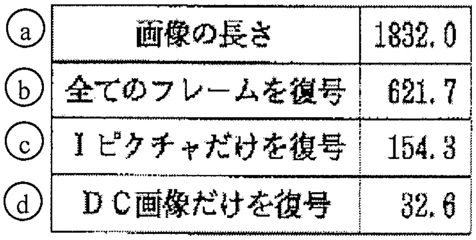
[0038]

In actual practice, when all of the frames of an approximately 30-minute MPEG-1 moving picture (the GOP being of the type shown in Figure 5) are decoded, Table 1 shows the processing time when only the DC pictures of the I-pictures are decoded. It can be seen that decoding can be performed in approximately 1/20 of the processing time required to decode all of the frames.

[0039]

Table 1

Table 1. Comparison of Decoding Time



ⓐ	画像の長さ	1832.0
ⓑ	全てのフレームを復号	621.7
ⓒ	Ｉピクチャだけを復号	154.3
ⓓ	ＤＣ画像だけを復号	32.6

Key:		
	a	Picture length
	b	Decoding of all frames
	c	Decoding of I-pictures only
	d	Decoding of DC pictures only

[0040]

In addition, Figure 9 shows an example of a DC picture and its source picture.

[0041]

In general, the degree of similarity that is used to compare frames can involve either a comparison of pixel values or a comparison of color histograms. The use of comparison by means of color histograms as the degree of similarity is convenient due to the fact that there is little influence due to movement of the camera or the photographic subject; however, it does not include half-screens or spatial information at all, which is problematic when completely different pictures have identical color histograms. There have been various attempts to include spatial information in a comparison by means of color histograms, but they all require somewhat complicated processing.

[0042]

With the present invention, $D_{histarea}$ of Equation (2) is used as the distance according to the color histogram, D_{pixsum} of Equation (3) is used as the distance according to the pixel value, and these two values are combined to calculate the degree of similarity; thus, a value for degree of similarity that encompasses spatial information can be obtained without sacrificing the simplicity of the process.

[0043]

Equation 2

$$D_{histarea} = 1 - \sum_{n=0}^{N_{bin}} \min(I_n, J_n) \Big/ \sum_{i}^{N_{bin}} I_n \qquad (2)$$

[0044]

Equation 3

$$D_{pixsum} = \frac{1}{H \cdot W} \sum_{i=0}^{H} \sum_{j=0}^{W} |a_{ij} - b_{ij}| \qquad (3)$$

[0045]

Here, the technique for calculating the degree of similarity from the two values uses a simplified fuzzy logic. By using fuzzy logic, the relationship between the degree of similarity of pictures, the distance according to the color histogram and the distance according to the pixel value can be described without being rigorously formulated. Furthermore, deductions based on simplified fuzzy logic are easy and can be executed quickly.

[0046]

The fuzzy logic rules used at this time are shown in the following Equation 4.

[0047]

Equation 4

$$\begin{array}{llccl} \text{rule } i: & \textsc{If} & D_{histarea} & \text{is} & A_a \\ & & D_{pixsum} & \text{is} & B_b \\ & \textsc{Then} & & s_i = c_i & \\ & & \multicolumn{3}{l}{a, b = \text{small}, \text{medium}, \text{large}} \\ & & \multicolumn{3}{l}{i = 1, 2, \cdots, 9} \end{array} \qquad (4)$$

[0048]

Here, i is the rule number, I is the number of rules, and c_i is a real value that represents the post-condition having values of [0.1]. In addition, A_a and B_b respectively are membership coefficients of the characteristic value, with the three membership coefficients "small," "medium," "large," as shown in Figure 10, being set for each characteristic value.

[0049]

The goodness of fit with respect to this rule is obtained with Equation (5); then, the final inference result, that is, the degree of similarity s between pictures, is determined with Equation (6). s has the value of [0,1].

[0050]

Equation 5

$$w_i = \mu_{A_a}(D_{histarea}) \cdot \mu_{B_b}(D_{pixsaura}) \qquad (5)$$

[0051]

Equation 6

$$s = \sum_{i=1}^{l} w_i c_i \Big/ \sum_{i=1}^{l} w_i \qquad (6)$$

[0052]

The detection of a shot is performed by detecting a cut point between shots.

[0053]

The detection of a cut point involves only the detection of a point for which the correlation between frames is low, but with the present invention, the correlation is not checked by directly comparing the characteristic amount of frames; instead, the correlation is checked based on the MPEG-1 encoding condition, thus, detecting the cut point. In other words MPEG-1 performs compression by means of a prediction based on the correlation between frames; conversely, the correlation between frames is checked by checking the manner in which the prediction is performed.

[0054]

By using the MPEG-1 encoding information without re-checking the characteristic amount of the frame, the number of calculations can be decreased; in addition, because it is not necessary to decode all of the information, the process can be performed quickly.

[0055]

As the processing procedure, first the cut point is detected based on the condition of the reference for the B-pictures,, then the reference for the P-pictures and the I-picture change are checked and confirmed. An example of a GOP for which N =15 and M = 3 will be explained using Figure 11.

[0056]

With a B-picture, the previous and following I- and P-pictures are referenced; in other words, typically four types of macroblocks MB, (IMB), (FMB), (BMB), and (BiMB), exist in a B-picture. With the IMB, referencing is not performed; the FMB references forward; the BMB, backward; the InMB [sic; BiMB], in both directions. The reference conditions are as shown in Figure 12.

[0057]

When the correlation among shots, i.e., between frames, is high, the number of previous and future references is almost equal; however, when a cut point exists between the picture and the referenced frame, the dependence is biased significantly toward the past or the future, and a bias is generated in the macroblock. This situation is shown in Figure 12(b), (c) and (d). However, Figure 12 represents an extreme case; in actual practice, it cannot be deemed that there is no reference that exceeds a cut point at all.

[0058]

Thus, it is clear that it is possible to determine the condition of the references to the previous and subsequent frames of a B-picture based on the structure of the B-picture's macroblock type. This is defined as the degree of dependence "relat" for a B-picture, as in Equation (7).

[0059]

Equation 7

$$relat = \frac{N_F - N_B}{N_{Bi}} \qquad (7)$$

[0060]

N_F, N_B and N_{Bi} are "respectively" the number of FMB, BMB and BiMB contained in the B-picture.

[0061]

The absolute value of relat increases as the difference between N_F and N_B increases and N_{Bi} decreases, indicating that the bias of the reference is larger.

[0062]

With a GOP as shown in Figure 11, if one focuses on two P-pictures (or I-pictures) and two B-pictures sandwiched therebetween (for example, f_7, f_8, f_9, f_{10} [sic; f_1, f_8, f_{10}, f_{11}]) and represents these as P_1 B_2 B_3 P_4, all of the cut points must be expressed with any one of the following patterns: P_1 |B_2 B_3 P_4, P_1 B_2 | B_3 B_4, P_1 B_2 B_3 | P_4 (with the "|" indicating a cut point). These correspond to Figure 12(b), (c) and (d). At this time, as is clear from Figure 12, in any of the cases, a reference bias occurs with both B_1 and B_2, and the absolute value of the degree of dependence becomes large.

[0063]

Accordingly, if Equation 8 is fulfilled, it is judged that a cut point exists in one of the patterns P_1 |B_2 B_3 P_4, P_1 B_2 | B_3 B_4, P_1 B_2 B_3 | P_4.

[0064]

Equation 8

$$|relat_{B_1} \cdot relat_{B_2}| > threshold_B \qquad (8)$$

[0065]

Next, a judgment is made as to which pattern it is: P_1 |B_2 B_3 P_4, P_1 B_2 | B_3 B_4, or P_1 B_2 B_3 | P_4, and the exact cut is determined. As is clear from Equation 7, the value of relat is positive when there are many prior references, and negative when there are many future references. Utilizing this fact, the cut point can be determined as in Equation (9).

[0066]

Equation 9

$$\begin{array}{lll} \mathrm{P_1|B_2B_3P_4} & \text{if} & relat_{B_1} < 0, \quad relat_{B_1} < 0 \\ \mathrm{P_1B_2|B_3P_4} & \text{if} & relat_{B_1} > 0, \quad relat_{B_1} < 0 \\ \mathrm{P_1B_2B_3|P_4} & \text{if} & relat_{B_1} > 0, \quad relat_{B_1} > 0 \end{array} \qquad (9)$$

[0067]

Thus, a cut point can be detected based on the B-picture reference information.

[0068]

When the detection is only based on the B-picture references, if there is noise or a momentary variation in the screen, such as an object crossing in front of the camera, a detection error may occur. This is thought to be because the distance between the B-picture and the picture being referenced is small. Therefore, the result is verified using the reference information for a P-picture, which references a picture that is farther away from the B-picture.

[0069]

When a cut point exists between the P-picture and the I- or P-picture being referenced, it is almost impossible to perform the reference, so that it is expected that the number of IMB will increase. Accordingly, when Equation (10) is fulfilled, a cut point obtained from a B-picture in the interval is considered a detection error and is deleted.

[0070]

Equation 10

$$\frac{N_I}{N} < threshold_P \qquad (10)$$

[0071]

N_I is the number of IMB contained in the P-picture, and N is total number of MB in the P-picture.

[0072]

A confirmation of the result is performed using I-pictures farther from the P-picture.

[0073]

Furthermore, with a GOP of the type shown in Figure 11, it is not possible to use a P-picture to confirm the result with respect to the cut point based on the two B-pictures (f_1, f_2) in front of the I-picture f_3, since the I-picture f_3 does not perform a reference. It is necessary here, as well, to detect detection errors by confirming the result using I-pictures.

[0074]

The color histogram distance $D_{histarea}$ (Equation (2)) is checked with respect to the DC pictures of the I-picture (f_3) and an I-picture for the GOP immediately prior, and when the following Equation (11) is fulfilled, a cut point obtained from a B-picture in the interval is considered a detection error and is deleted.

[0075]
Equation 11

$$D_{histarea} < threshold_{I} \qquad (1\,1)$$

[0076]

Next, the selection of the representative frame of a shot will be explained. Compared to the overall moving picture, a shot is a short unit; however, for example, a 5-second shot is a collection of 150 frames (in the case of 30 f_{ps}), and it is difficult to perform processes such as comparison, display, and detection of the characteristic amount in this situation. Therefore, when a shot is handled, typically, a frame from said shot is selected as representative, and this representative frame is used to perform processes such as comparison and display.

[0077]

With the present invention, when shots are integrated and a scene is extracted, it is used to determine the degree of similarity between shots. It also is used to show the content of the shot in a simple manner when the moving picture structure that is analyzed is presented to the user. Therefore, a frame that best represents the content of the shot must be selected.

[0078]

With conventional research dealing with a shot, the leading frame or central frame of the shot is often used automatically as the representative frame. However, it is difficult to say that the frame thus selected well represents the content of the shot. Accordingly, with the present invention, the frame that is closest to the average of the frames contained in the shot is selected as the representative frame.

[0079]

Furthermore, in addition to shots without motion, there are, for example:
(1) those with camera motion in the middle of the shot;
(2) those for which the camera or an object within the picture continues to move; and
(3) those with a very large amount of motion.
With such shots, it is difficult to represent the entire shot with a single frame, and there is the danger that useful information will be lost. Therefore, these types of shots are represented by multiple representative frames.

[0080]

Furthermore, by selecting multiple representative frames, the influence of misdetection of the cut position can be minimized. In other words, when what should be multiple shots ends up being compiled into one due to a detection mistake and when only one representative frame is selected, then, essentially, the information for only one shot can be used. On the other hand, if multiple selections are made based on the content, they can be utilized when extracting a scene without discarding the information of the various shots.

[0081]

To select the required minimum number of representative frames, first, the frames in the shot are clustered. For each of the resulting clusters, the frame closest to the average is selected as the representative frame of the shot.

[0082]

However, if all of the frames in a shot are made candidates for the representative frame, there is the possibility that the amount of processing will increase for a longer shot. Therefore, only I-pictures are used as candidates; this allows the amount of processing involved in the selection, as well as the amount of processing involved in decoding from the source moving picture to be reduced. Furthermore, typically, with MPEG-1, the quantization characteristics of the I-, P- and B-pictures are changed to improve encoding efficiency, so that this is also advantageous, since the I-pictures often have the best quality.

[0083]

Furthermore, the I-pictures are not completely decoded; as explained above, the DC pictures are used to achieve a reduction in the amount of processing during decoding.

[0084]

The actual processing procedure is as follows (Figure 13). When there are no I-pictures in the shot, i.e., when the shot is extremely short, the representative frame is the first P-picture appearing in the shot, and when there is none, the first B-picture. With an extremely short shot, it can be said that there is almost no change within the shot, so that an automatic process of this type is sufficient.

[0085]

(1) The I-pictures contained in the shot undergo simplified decoding, and DC pictures are extracted.

[0086]

(2) DC pictures for which there is little motion in the shots are initially clustered. To judge whether there is little motion, the number of IMB contained in the B- or P-pictures is checked (Equation (12)).

[0087]

Equation 12

$$N_I < threshold_{move} \tag{12}$$

[0088]

(3) Clustering is performed based on the aforementioned initial clusters, and the I-pictures in the shot are classified into a number of clusters. The clustering is performed using the group average method, and $D_{histarea}$ (Equation (2)) is used as the distance between elements. Clustering is performed until the shortest distance between clusters exceeds a threshold value.

[0089]

(4) After clustering is completed, a representative frame is selected from each cluster. First, the DC pictures of the I-pictures in the clusters are averaged to create an average DC picture (Equation (13)).

[0090]

Equation 13

$$\begin{aligned} \overline{Y_{ij}} &= \frac{1}{N}\sum_{n=1}^{N} Y_{ij}^{n} \\ \overline{Cb_{ij}} &= \frac{1}{N}\sum_{n=1}^{N} Cb_{ij}^{n} \\ \overline{Cr_{ij}} &= \frac{1}{N}\sum_{n=1}^{N} Cr_{ij}^{n} \end{aligned} \tag{13}$$

Here, N is the number of DC pictures, $\overline{Y_{ij}}$, $\overline{Cr_{ij}}$, $\overline{Cb_{ij}}$ and are the average luminance and color difference signals for the coordinates (i, j) of the DC pictures, Y_{ij}^{n}, Cb_{ij}^{n}, and Cr_{ij}^{n} and indicate the average luminance and color difference signals for the coordinates (i, j) of the nth DC picture, DC_n.

[0091]

(5) The I-picture I_k having the DC picture DC_k for which the distance D_{pixsum} (Equation (3)) to the average DC picture is the shortest is made the representative frame.

[0092]

Thus, the representative frame for a shot is selected.

[0093]

During actual processing, selection of the representative frame is performed in parallel with detection of the cut points to efficiently obtain the DC pictures of the I-pictures and the macroblock information for the P- and B-pictures.

[0094]

For example, with a scene of a conversation or the like, often the speakers are alternately photographed, so that the same type of shot is repeated; thus, a number of similar shots often are contained in one scene. With attention to this predisposition, the shots are integrated, and the scene is extracted.

[0095]

The degree of similarity between shots utilizes the degree of similarity s (Equation (6)) between the representative frames of the respective shots. However, one shot may have several representative frames, so that the degree of similarity for the combination of all of the representative shots is checked, and the maximum value of the results thereof is taken as the degree of similarity of the shot (Figure 14).

[0096]

As shown in Figure 15, the simplest method for extracting a scene from a shot is a method in which, when there are similar shots (which have an extremely high degree of similarity), everything in between is considered the same scene.

[0097]

However, in this case, there are the following problems:

(1) The threshold value setting used in determining whether shots are similar has a significant impact on the result.

(2) There is a lack of flexibility, in that, when there is no group of shots for which the degree of similarity is extremely high, but there are many groups of shots with a moderate degree of similarity, they cannot be considered to be the same scene.

[0098]

Therefore, a degree of connection connect$_{n,n+1}$, which represents the degree to which a shot shot$_n$ and a shot shot$_{n+1}$ are connected (i.e., belong to the same scene) is defined as in Equation (14), and a scene is extracted using this degree of connection.

[0099]

Equation 14

$$connect_{n,n+1} = 1 - \prod_{i=n-N+1}^{n} \prod_{j=n+1}^{i+N} (1 - s_{ij}) \qquad (1\,4)$$

[0100]

Here, N indicates the range of the shots being compared, and s_{ij} is the degree of similarity between shot$_i$ and shot$_j$.

[0101]

Thus, the degree of connection connect$_{n,n+\ 1}$ is determined not only from the shot shot$_n$ and the shot shot$_{n+1}$, but also from the degree of similarity s_{ij} between all of the shots in the vicinity thereof. For example, in Figure 16, connect$_{3,4}$ is determined using not only the degree of similarity s_{34} of shot$_3$ and shot$_4$, but also the degree of similarity s_{25} of shot$_2$ and shot$_5$. This is so that even if shot$_3$ and shot$_4$ e.g., are in no way similar, shot shot$_3$ and shot shot$_4$ can be considered to belong to the same scene if shot$_2$ and shot$_5$ are similar.

[0102]

However, the probability that shots that are extremely separated temporally belong to the same scene is low; instead, there is a possibility that shots with a high degree of similarity may exist regardless of whether they belong to the same scene, and to prevent, insofar as possible, their going undetected as a result of such a situation, the range of the shots that are compared is restricted to N.

[0103]

An example of the change in the degree of connection $connect_{n,\,n+1}$ obtained with Equation (13) is shown in Figure 17.

[0104]

Based on such changes in the degree of connection, a change of scene can be determined, and scenes can be extracted. Here, when the difference between a peak and a valley of the change is greater than a threshold value $threshold_{SCENE}$, the cut point that has the degree of connection that is the valley thereof is made the point where the scene changes.

[0105]

Effect of the invention

By means of the present invention, a provided moving picture is hybrid-encoded and is divided into a hierarchical structure; shots are integrated based on the degree of similarity between the divided shots to extract scenes; so that the effect is that the scenes can be extracted quickly and accurately. Therefore, it is relatively easy to extract a shot or a frame from a scene, and, thus, the effect is that a scene, shot, or cut [point] can be extracted from a moving picture quickly and accurately.

[0106]

All of the aforementioned processes can be automated with existing hardware or appropriately assembled hardware, so that a prescribed scene, shot, or cut [point] can be provided automatically by means of an appropriate command input.

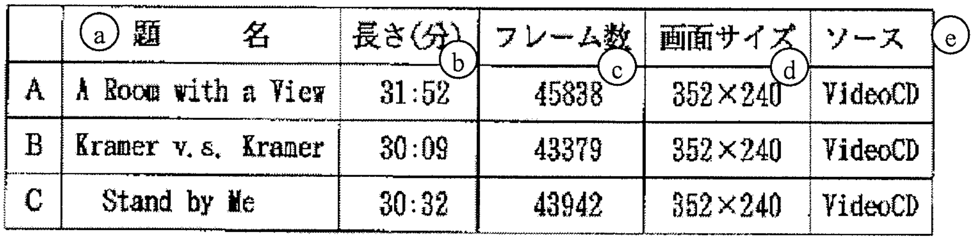
[0107]

The cut detection results obtained by experimentation using the moving pictures in Table 2 are shown in Table 3.

[0108]

Table 2

Table 2. Moving Images Used in Evaluation



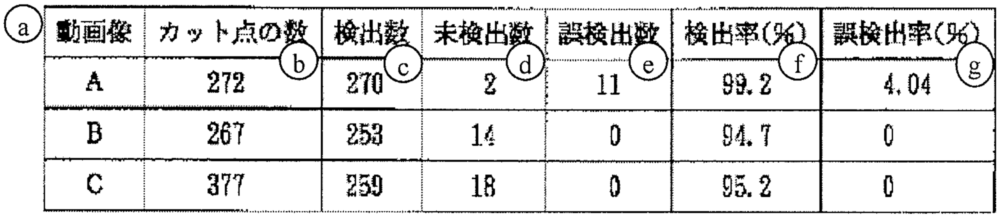
	ⓐ 題　　名	長さ(分) ⓑ	フレーム数 ⓒ	画面サイズ ⓓ	ソース ⓔ
A	A Room with a View	31:52	45838	352×240	VideoCD
B	Kramer v.s. Kramer	30:09	43379	352×240	VideoCD
C	Stand by Me	30:32	43942	352×240	VideoCD

Key:	a	Name
	b	Length (minutes)
	c	Number of frames
	d	Screen size
	e	Source

[0109]

Table 3

Table 3. Cut point detection results



動画像	カット点の数	検出数	未検出数	誤検出数	検出率(%)	誤検出率(%)
A	272	270	2	11	99.2	4.04
B	267	253	14	0	94.7	0
C	377	259	18	0	95.2	0

Key:	a	Moving picture
	b	Number of cut points
	c	Number detected
	d	Number undetected
	e	Number of detection errors
	f	Detection rate (%)
	g	Detection error rate (%)

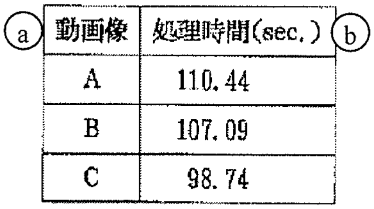
[0110]

In addition, the cut point detection processing time is shown in Table 4.

[0111]

Table 4

Table 4. Cut Point Detection Processing Time



動画像	処理時間(sec.)
A	110.44
B	107.09
C	98.74

Key:	a	Moving picture
	b	Processing time

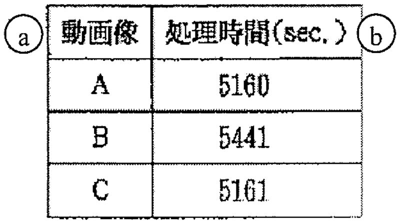
[0112]

Furthermore, the processing time for cut point detection by means of a simple algorithm is shown in Table 5.

[0113]

Table 5

Table 5. Cut point Detection Processing Time by Means of Simple Algorithm



動画像	処理時間(sec.)
A	5160
B	5441
C	5161

Key: a Moving picture
b Processing time

[0114]

Next, scenes were extracted using the shots obtained with the cut point detection results. The change in the degree of connection is shown in Figures 18, 19 and 20.

[0115]

The results when scenes were extracted based on the degrees of connection in Figures 18, 19 and 20 are shown in Table 6. The threshold value $\text{threshold}_{\text{SCENE}}$ for detection of the point where a scene changes was 0.3.

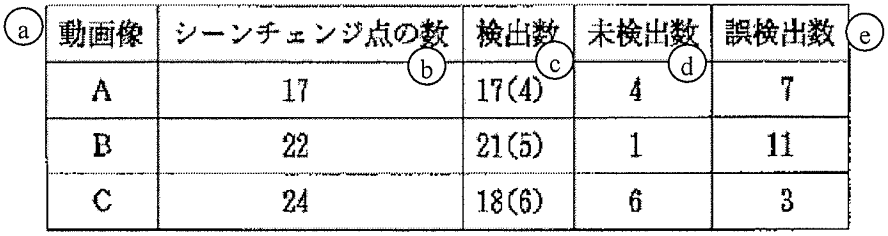
[0116]

From Table 6, 75% or more of the scene change points were detected for all of the moving pictures, which is sufficiently practical considering the rapidity of the technique and the fact that the present technique does not use semantic analysis or information. In addition, approximately 1/4 to 1/3 of the number detected were one shot prior to or one shot after the actual scene change point. This is due to a defect in the algorithm in that, when a shot with a high degree of similarity to a shot that is adjacent to the scene change point does not exist in the scene to which it should belong, the resulting degree of connection is low.

[0117]

Table 6

Table 6. Scene Change Point Detection Results



動画像	シーンチェンジ点の数	検出数	未検出数	誤検出数
A	17	17(4)	4	7
B	22	21(5)	1	11
C	24	18(6)	6	3

Key:		
	a	Moving picture
	b	Number of scene change points
	c	Number detected
	d	Number undetected
	e	Number of detection errors

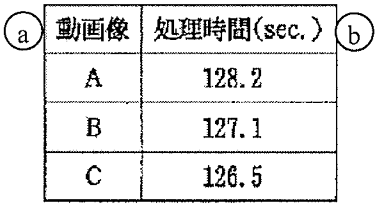
[0118]

The overall processing times for the structural analysis process are shown in Table 7; it can be seen that they are sufficiently quick.

[0119]

Table 7

Table 7. Structural Analysis Processing Time



動画像	処理時間(sec.)
A	128.2
B	127.1
C	126.5

Key:		
	a	Moving picture
	b	Processing time

Brief description of the figures

Figure 1 is the MPEG-1 encoding/decoding system of this present invention.

Figure 2 is a block diagram of said MPEG-1 video encoder.

Figure 3 is a diagram of said MPEG-1 video decoder.

Figure 4 is a diagram of an example of GOP of same.

Figure 5 is a diagram showing an alignment of a source picture and the screens in the stream of same.

Figure 6 is a diagram of the hierarchical structure of said MPEG-1.

Figure 7 is a diagram of the flow of the structural analysis process of same.

Figure 8 is a diagram of the creation of a DC picture of same.

Figure 9 is an example of a DC picture of same.

Figure 10 is a shape diagram of the membership coefficient used in the fuzzy inference that obtains the degree of similarity for same.

Figure 11 is a an example of a GOP of same.

Figure 12 (a) is normal reference diagram; (b) is a diagram showing an example wherein a cut point exists between a past P-picture and [the B-pictures]; (c) is a diagram showing an example wherein a cut point exists between B-pictures; (d) is a diagram showing an example wherein a cut point exists between a future P-picture and [the B-pictures].

Figure 13 is a diagram showing the steps of the algorithm for selection of a reference frame: (a) is a diagram of the extraction of DC pictures; (b) is a diagram of the determination of the initial clusters; (c) is a diagram of the [final] clustering; (d) is a diagram of creation of average pictures from the DC pictures in the clusters; (e) is a diagram of the selection of frames closest to the average pictures as reference frames.

Figure 14 is a diagram showing the degree of similarity between shots for same.

Figure 15 is a diagram of simple scene extraction for same.

Figure 16 is a diagram showing the degree of connection when N = 3 for same.

Figure 17 is a diagram showing an example of the change in the degree of connection connect $_{n,n+1}$ of same.

Figure 18 is a diagram showing an example of the degree of connection for moving picture A of same.

Figure 19 is a diagram showing an example of the degree of connection for moving picture B of same.

Figure 20 is a diagram showing an example of the degree of connection for moving picture C of same.

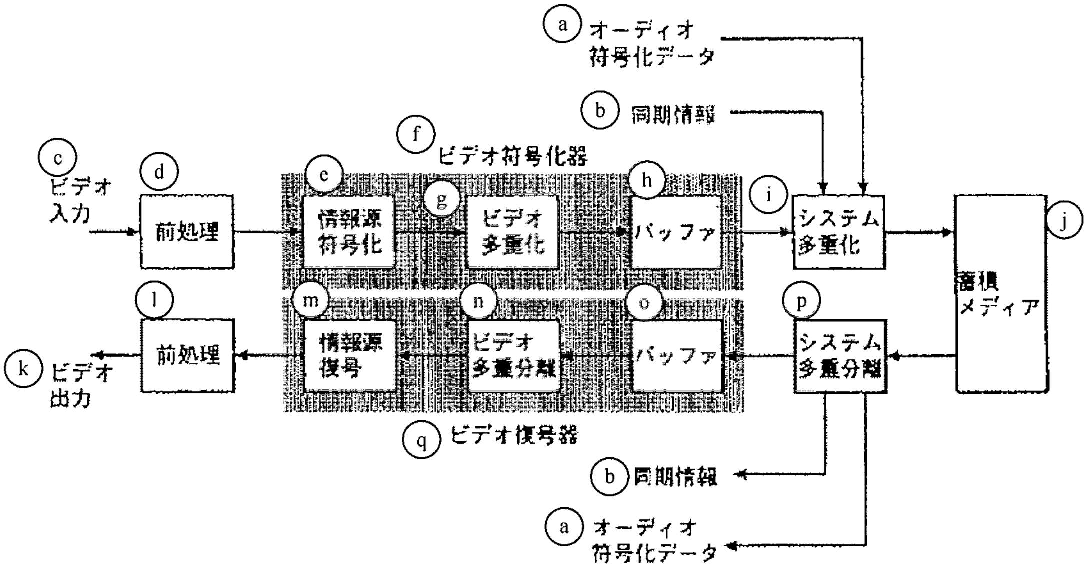


Figure 1

Key:	a	Audio encoded data
	b	Synchronization information
	c	Video input
	d	Pre-processing
	e	Information source encoding
	f	Video encoder
	g	Video multiplexing
	h	Buffer
	i	System multiplexing
	j	Storage medium
	k	Video output
	l	Pre-processing
	m	Information source decoding
	n	Video demultiplexing
	o	Buffer
	p	System demultiplexing
	q	Video decoder

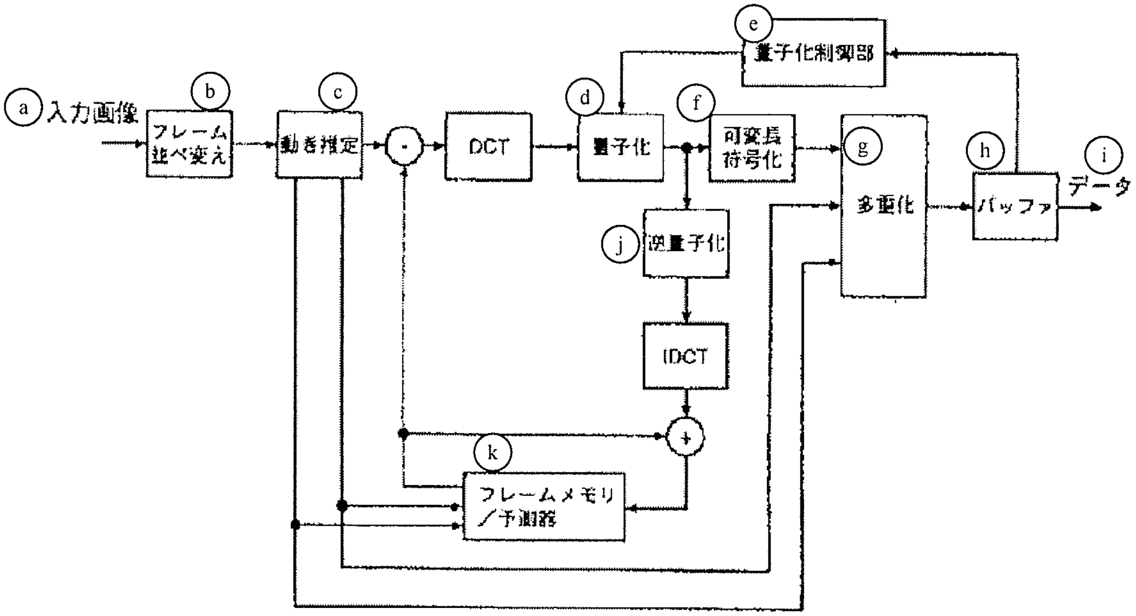


Figure 2

Key:	a	Input picture
	b	Frame alignment
	c	Motion estimation
	d	Quantization
	e	Quantization control unit
	f	Variable-length encoding
	g	Multiplexing
	h	Buffer
	i	Data
	j	Reverse quantization
	k	Frame memory/predictor

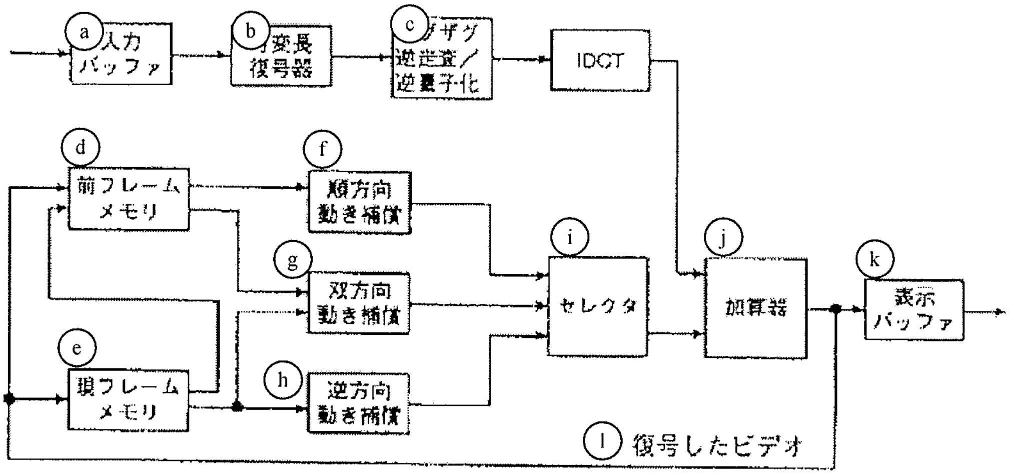


Figure 3

Key:	a	Input buffer
	b	Variable-length decoder
	c	Zigzag reverse scanning/reverse quantization
	d	Previous frame memory
	e	Current frame memory
	f	Forward motion compensation
	g	Bidirectional motion compensation
	h	Backward motion compensation
	i	Selector
	j	Adder
	k	Display buffer
	l	Decoded video

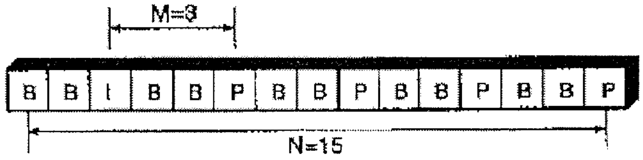


Figure 4

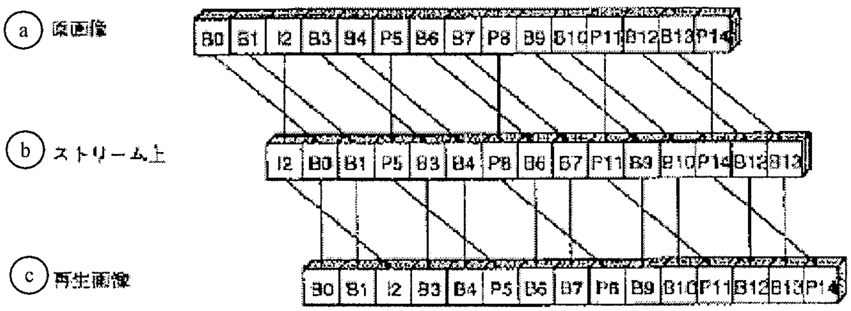

a 原画像
B0 B1 I2 B3 B4 P5 B6 B7 P8 B9 B10 P11 B12 B13 P14
b ストリーム上
I2 B0 B1 P5 B3 B4 P8 B6 B7 P11 B9 B10 P14 B12 B13
c 再生画像
B0 B1 I2 B3 B4 P5 B6 B7 P8 B9 B10 P11 B12 B13 P14

Figure 5

Key:		
	a	Source picture
	b	In stream
	c	Reproduced picture

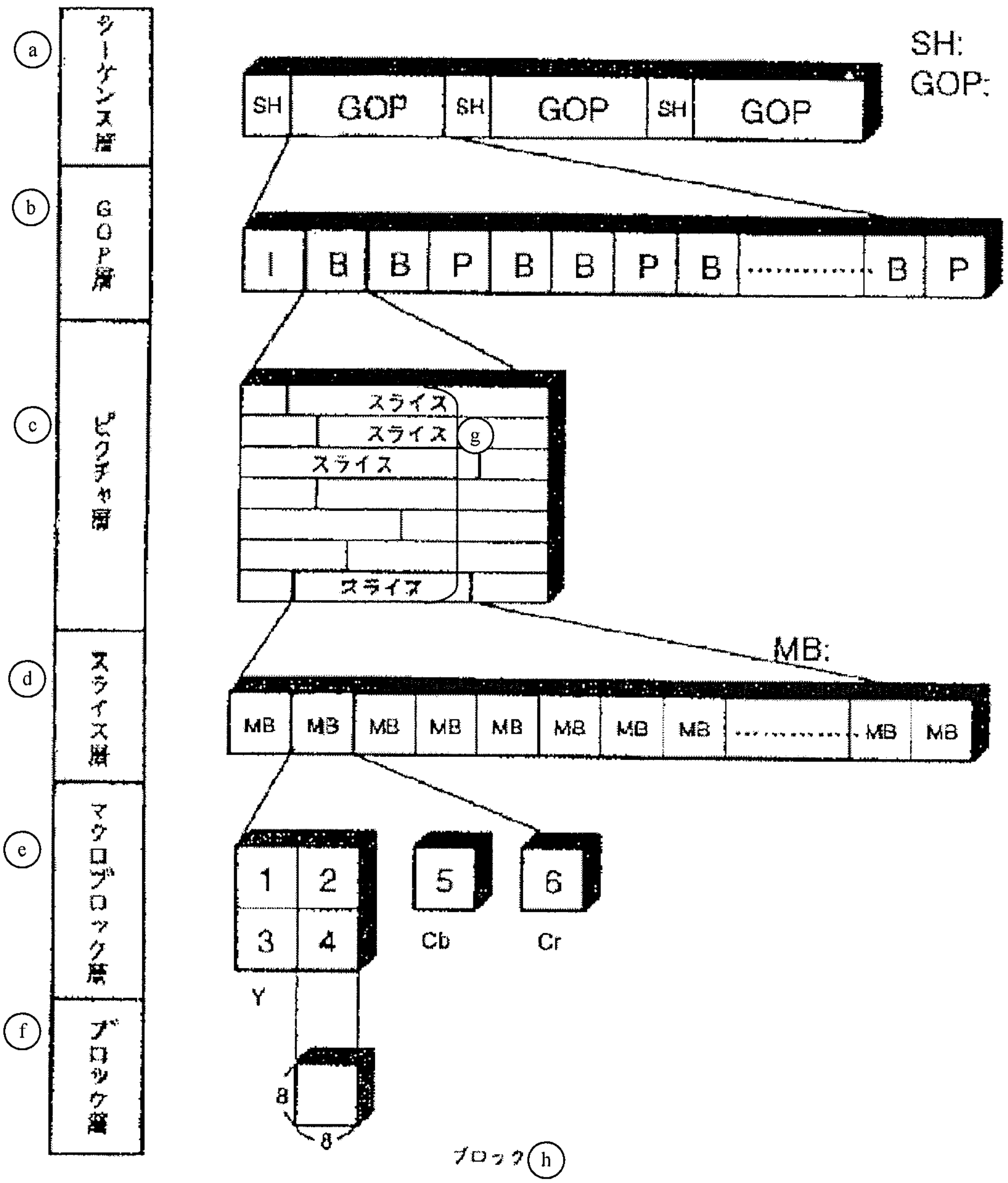


Figure 6

Key:

a	Sequence layer
b	GOP layer
c	Picture layer
d	Slice layer
e	Macroblock layer
f	Block layer
g	Slice
h	Block

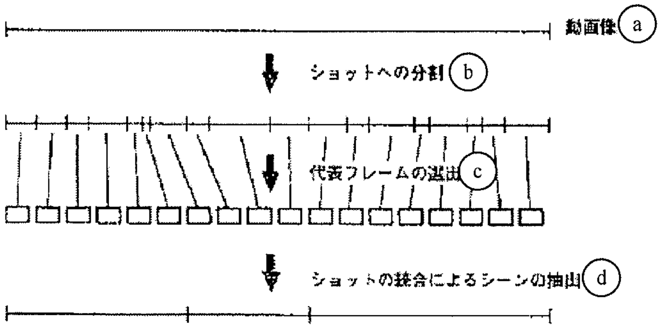


Figure 7

Key: a Moving picture
b Separation into shots
c Selection of reference frames
d Extraction of scenes based on integration of shots

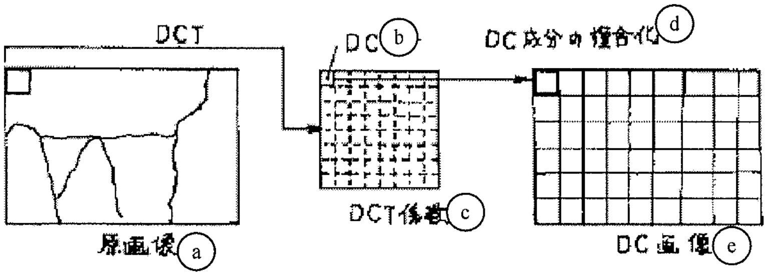


Figure 8

Key: a Source picture
b DC component
c DCT coefficient
d Decoding of DC component
e DC picture

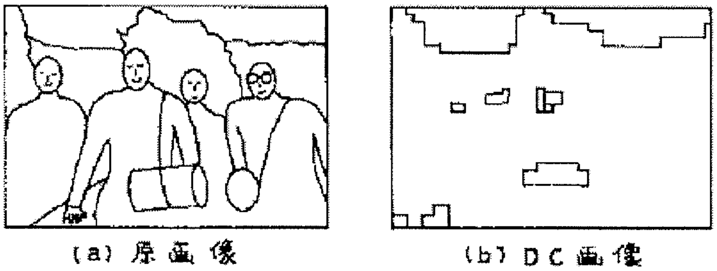


Figure 9

Key: (a) Source picture
(b) DC picture

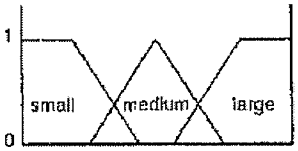


Figure 10

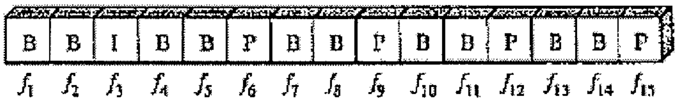


Figure 11

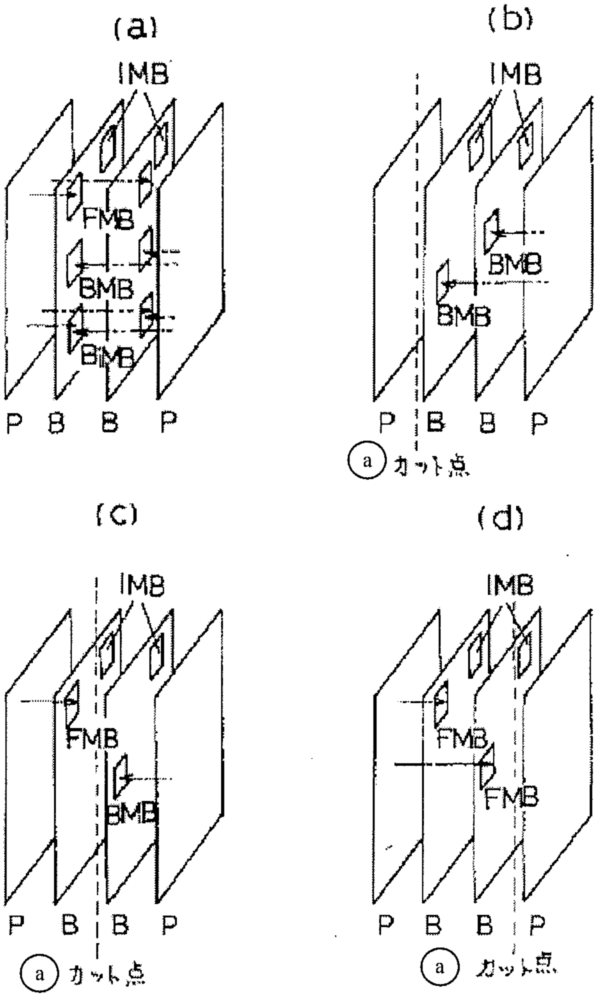


Figure 12

Key: a Cut point

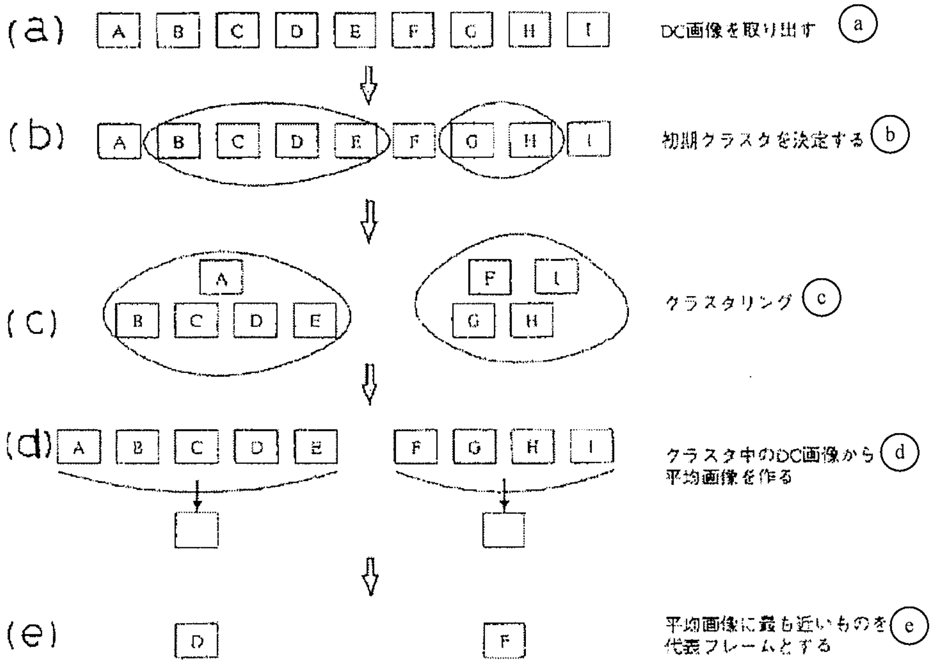


Figure 13

Key: a Extraction of DC picture
b Determination of initial clusters
c Clustering
d Creation of average pictures from DC pictures in clusters
e Selection of frames closest to average pictures as reference frames

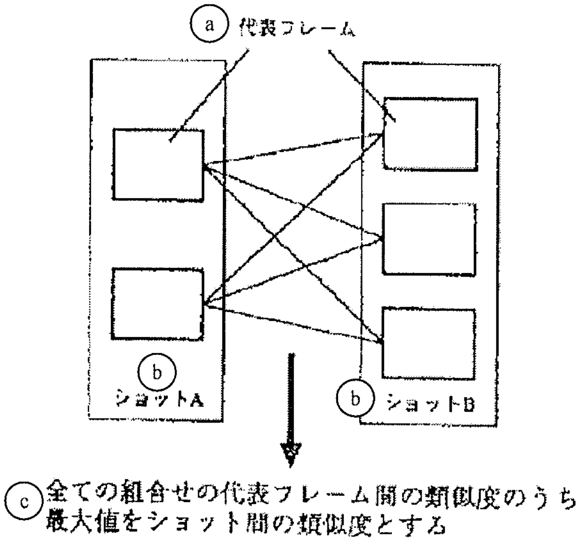

a 代表フレーム
b ショットA
b ショットB
c 全ての組合せの代表フレーム間の類似度のうち
最大値をショット間の類似度とする

Figure 14

Key: a Reference frame
b Shot
c The degree of similarity having the maximum value for all combinations of reference frames is made the degree of similarity between shots

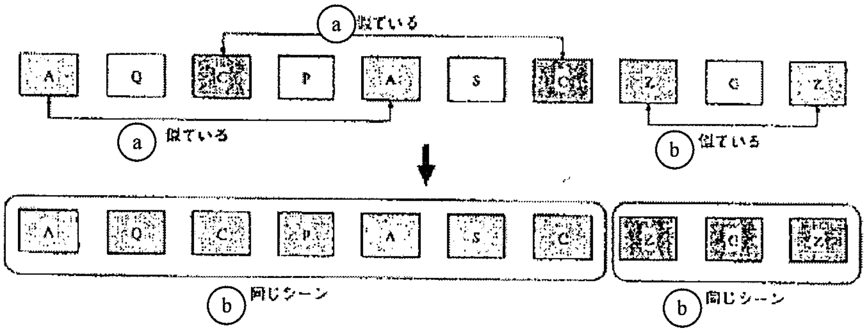


Figure 15

Key: a Similar
b Same scene

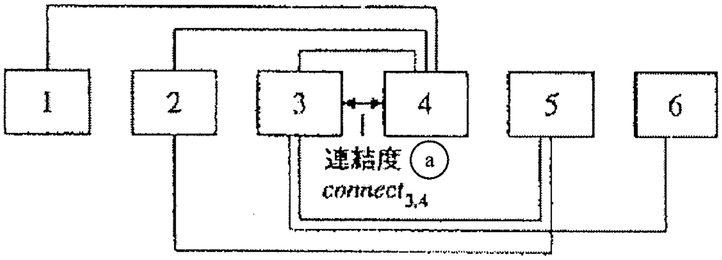


Figure 16

Key: a Degree of connection

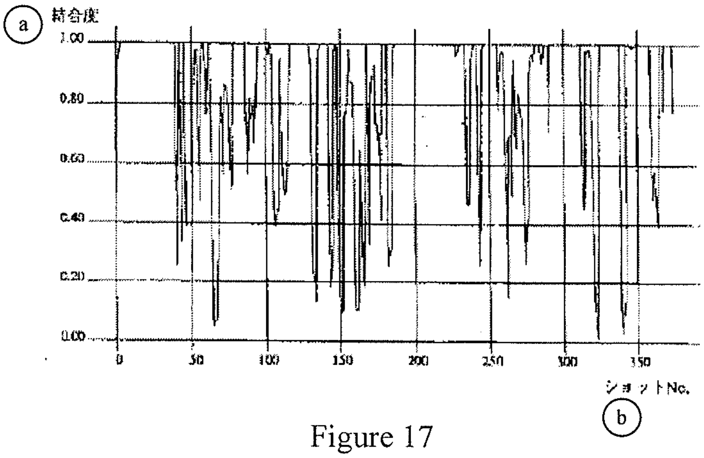


Figure 17

Key: a Degree of connection
b Shot

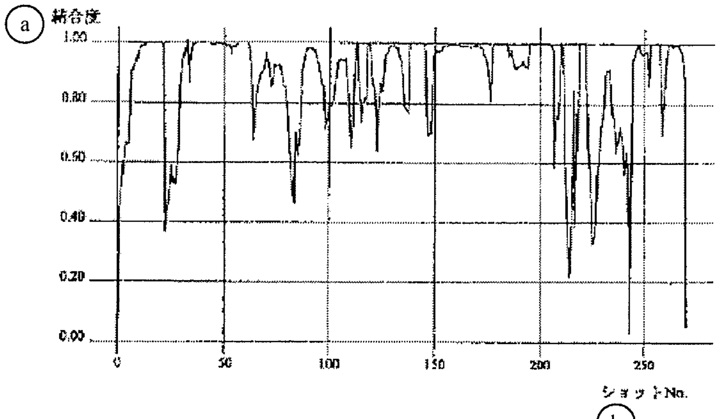


Figure 18

Key: a Degree of connection
b Shot

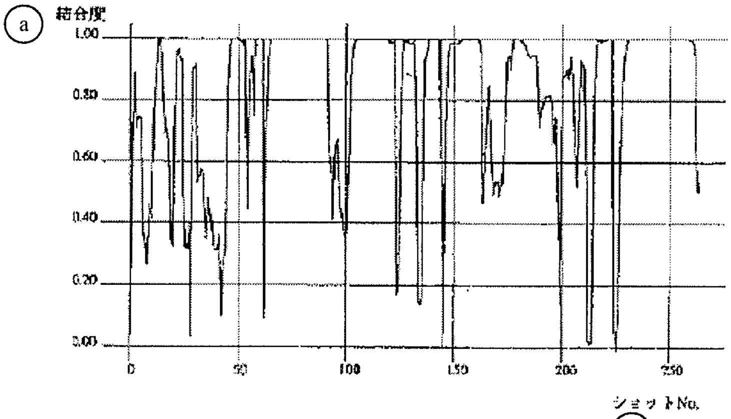


Figure 19

Key: a Degree of connection
b Shot

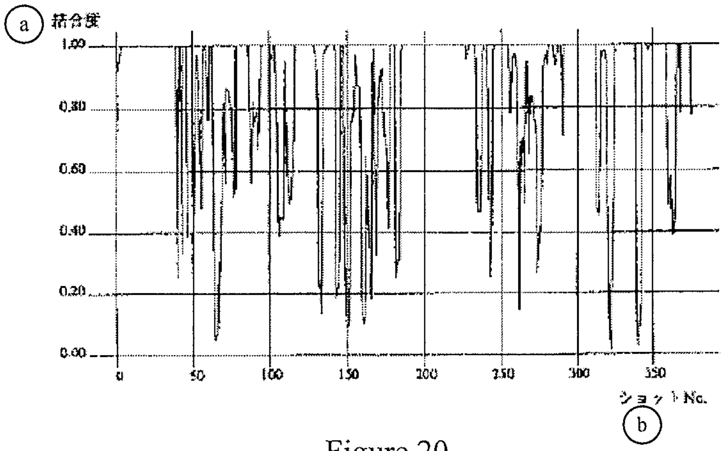


Figure 20

Key: a Degree of connection
b Shot